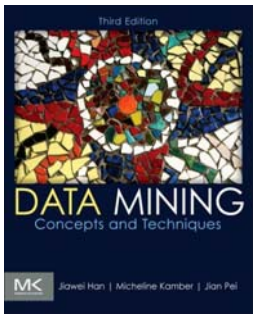
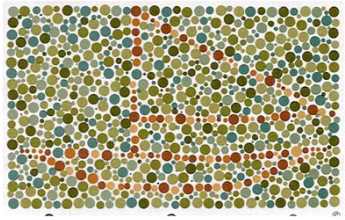
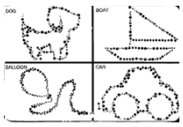
Data Mining



1

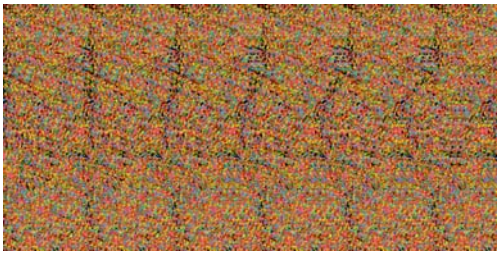
Vision Test?





2

Vision Test, or Data Mining?



3

Why taking data mining?

- You
- You background (UG/MS/PhD)
- Your expectation

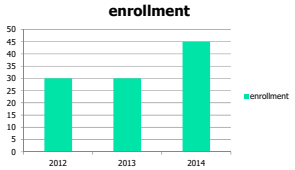
- My expectation:
- You have taken MTH 161/165, CSC 171/172
- Preferably you have taken CSC 242 or 262
- You love to work with data (or you think)

4

Why taking data mining?

5

Capping the class size



What happened since 2015?

6

Quote of the Day

Must do
Should do
Could do

7

What is data mining?

The *computerized* (sometimes iterative and interactive) process of discovering valid, novel, useful, and understandable patterns or models in

Massive databases

8

Big Data

9

The 5 Vs of Big Data

To get a better understanding of what Big Data is, it is often described using Five Vs:

10

What is data mining?

- Valid: generalize to the future
- Novel: what we don't know
- Useful: be able to take some action
- Understandable: leading to insight
- Iterative: takes multiple passes
- Interactive: *human* in the loop

11

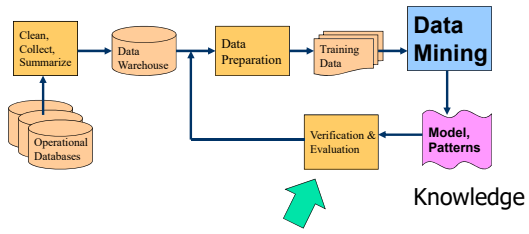
Data mining goals

- Prediction
  - Predict the value of a specific attribute based on the values of other attributes, e.g., medical diagnosis
  - "Opaque": ->What?
  - Approaches: classification, regression, outlier detection
- Description
  - Derive patterns (correlation, trends, trajectories) that summarize the underlying relationship between data, e.g., trending in twitter
  - "Transparent": ->Why?
  - Approaches: clustering, association rules, pattern discovery

12

KDD Process

Knowledge Discovery from Data



13

Data mining process

- Understand the application *domain*
  - Prior knowledge, user goals
- Create a target dataset
  - Select data, focus on subsets (*relevant to the task*)
- Data cleaning and transformation
  - Remove noise, outliers, missing values
  - Select features, reduce dimensions

14

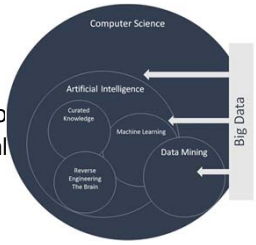
Data mining process

- Apply data mining algorithms
  - Associations, sequences, classification, clustering, etc.
- *Interpret*, evaluate and *visualize* patterns
  - What's new and interesting?
  - Iterate if needed
- Manage discovered knowledge
  - Close the loop!

15

Related fields

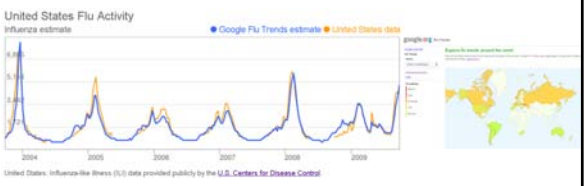
- Big data
- Data science
- Machine learning
- Statistics
- Databases and data wareho
- High performance and paral
- Visualization



16

Google Flu Trends

- Conjecture: When people get sick, some of them google the word "flu".
- Statistical analysis of the *crowd* allows the location and severity of the disease to be monitored, up to 2 weeks earlier than CDC using data from hospitals (Ginsberg et al., *Nature*, 2009).



17

Information Networks Are Everywhere

Social Networking Websites

Biological Network: Protein Interaction

They are all treated as Homogeneous Networks!

Research Collaboration Network

Product Recommendation Network via Emails

18

Heterogeneous Networks

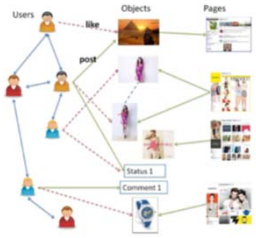


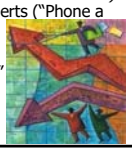
Figure 2: A heterogeneous network model for social media with 'like' function, using Facebook as an example. A blue bidirectional arrow is the friendship link. A red dashed arrow is a *like* action, while a green arrow denotes a *post* action, annotated with the time stamp when it was posted. The dashed blue line shows that the two photos are visually similar.

19

19

The Wisdom of Crowds

- Independent input from non-experts can sometimes accurately predict outcomes better than experts
- Examples:
  - In a 1906 fair, the mean of 787 guesses at the weight of a bull was within a pound
  - The crowd ("Ask the audience" in *Who Wants to Be a Millionaire?)* select the correct answer more often than the experts ("Phone a friend")
  - Maze navigation improves with crowd input
  - The stock market is a massive "Stupidity of crowds"



20 CSC 240/440 by Prof. Jiebo Luo

20

A Social Picture is Worth 1000 Votes

Ge Ma, Jiebo Luo, ICME 2013

- Image sentiment
  - Positive vs. negative campaigns
  - Positive vs. negative social media
  - Flattering vs. unflattering expressions

HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES — DEPARTMENT OF COMPUTER SCIENCE

21

2012: Calling the Swing States

- Acts like a prism to reveal the spectrum of opinions
- Competitive Vector Autoregressive Model

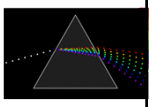


Table 2: Prediction for the swing states on Election Day

State	Official		AR		Flickr-AR		VAR		CVAR	
	Obama	Romney	Obama	Romney	Obama	Romney	Obama	Romney	Obama	Romney
CO	0.5239	0.4761	0.5073	0.4927	**0.5076**	**0.4924**	0.5059	0.4941	0.5049	0.4951
FL	0.5044	0.4956	0.4936	0.5064	0.4928	0.5072	0.4917	0.5083	**0.5018**	**0.4982**
IA	0.5287	0.4713	0.5139	0.4861	0.5132	0.4868	0.5153	0.4847	**0.5291**	**0.4709**
NC	0.4891	0.5109	0.4851	0.5149	0.4845	0.5155	**0.4866**	**0.5134**	0.4888	0.5512
NV	0.5335	0.4665	0.5144	0.4856	0.5144	0.4856	0.5165	0.4835	**0.5362**	**0.4638**
OH	0.5098	0.4902	0.5155	0.4845	0.5150	0.4850	**0.5105**	**0.4895**	0.5120	0.4880
VA	0.5157	0.4843	0.5022	0.4978	0.5019	0.4981	0.5013	0.4987	**0.5185**	**0.4815**
WI	0.5339	0.4661	0.5218	0.4782	0.5219	0.4781	**0.5285**	**0.4715**	0.5460	0.4540

22

Fine-Grained Analysis of the 2016 Election

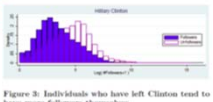


Figure 3: Individuals who have left Clinton tend to have more followers themselves.

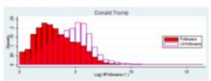


Figure 4: Individuals who have left Trump tend to have more followers themselves.

For both candidates, we find that it is those who themselves have more followers that are more likely to leave.

4.2 Gender

In this subsection, we analyze the gender composition of Clinton's and Trump's unfollowers.

sharply.* This makes our gender analysis of the Clinton unfollowers particularly interesting.

Figure 5: Compared with current followers, Clinton's unfollowers are more likely to be female.

Figure 6: Compared with current followers, Trump's unfollowers are more likely to be female. The effect is *less* pronounced than Clinton.

23

America Tweets China: Analysis of State and Individual Characteristics Regarding Attitudes towards China

Main Contributors: Yu Wang and Jiebo Luo, IEEE Big Data Conference, 2015

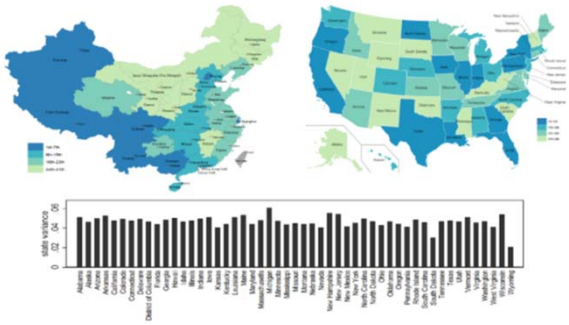


Figure 5: Friendliness and Variance. The top figure reports the friendliness index for each state. The bottom figure reports the variance index for each state.

24

This is also Data Science

25

Twitter Health

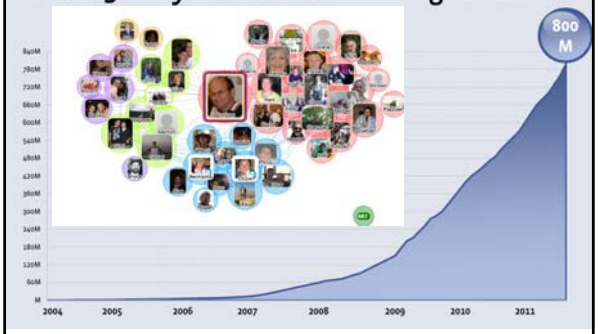
Adam Sadilek, Henry Kautz, and Vincent Silenzio, Predicting Disease Transmission from Geo-Tagged Micro-Blog Data, AAAI 2012

- You can explore health patterns with the web application at GermTracker

26

Courtesy of Dr. Rong Yan

Facebook reaches the milestone of ~~800 Million~~ 1 Billion ~~30-day active users on Aug. 2011~~ Oct. 2012



27

Courtesy of Dr. Rong Yan

28

Courtesy of Dr. Rong Yan

Social Media has a Pervasive Influence on Daily Life

Source: stories.facebook.com

29

Courtesy of Dr. Rong Yan

Mining Large-Scale Social Multi-Media

Understand users better and improve their social experience?

Connection	• Who should be suggested to you as new friends • How to measure and discover online communities
Communicate	• What are users talking about in their news feed • How to analyze user rating and reputation • User sentiment and emotion in blogsphere
Influence	• How does the memes influence user behavior • What are your viral marketing strategies
Identity	• What kinds of ads are relevant to a user • Can we suggest face tags for each photo • How to recommend the best images / videos

People Analytics

30

Courtesy of Dr. Rong Yan

The Pursuit of Happiness

- Status update could be an measure of happiness
  1. Representative (unbiased sampled population)
  2. Unobtrusive (based on naturalistic behavior)
  3. Computational (no human raters)
  4. Correlated (co-vary with other valid measures, e.g., non-naturalistic self-reports)
  5. Efficient (process millions of updates per day)

31

Courtesy of Dr. Rong Yan

Facebook Gross National Happiness

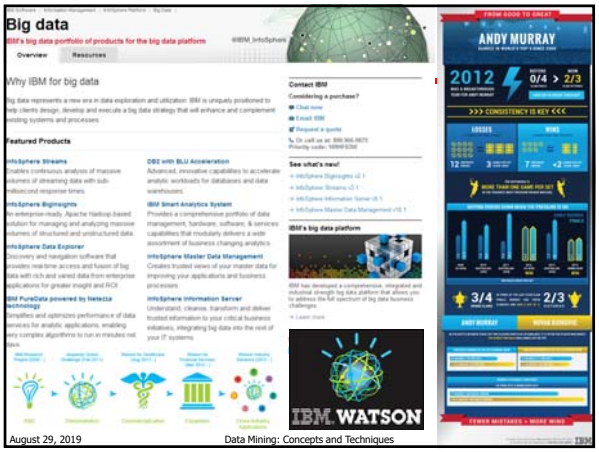
- Estimate each country's average happiness based on status updates [http://apps.facebook.com/gnh_index]

32

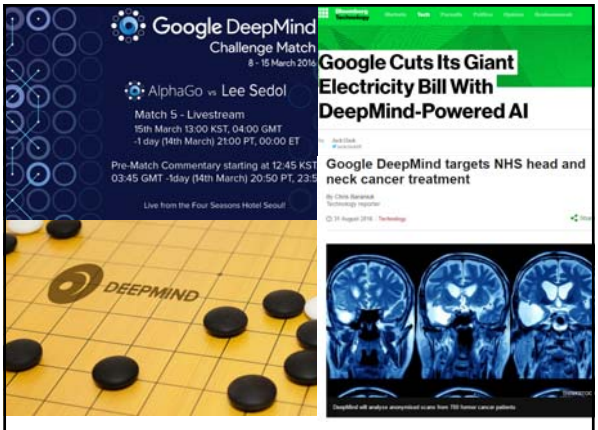
Facebook Mood Manipulation

- Facebook has been experimenting on users. A new paper in the *Proceedings of the National Academy of Sciences* (PNAS) reveals that Facebook intentionally manipulated the news feeds of almost 700,000 users in order to study "emotional contagion through social networks."
- Army Research Office + Cornell University
- Over the course of the study, it appears, the social network made some of us happier or sadder than we would otherwise have been. Now it's made all of us more mistrustful!

33



34

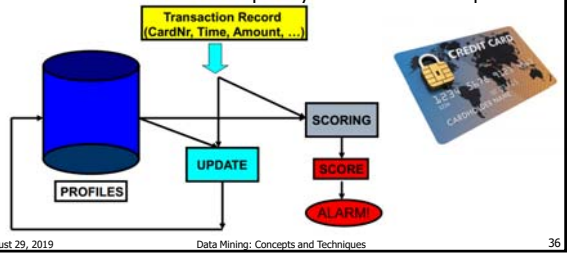


35

Credit Card Fraud Detection

Trouble with rule-based approaches

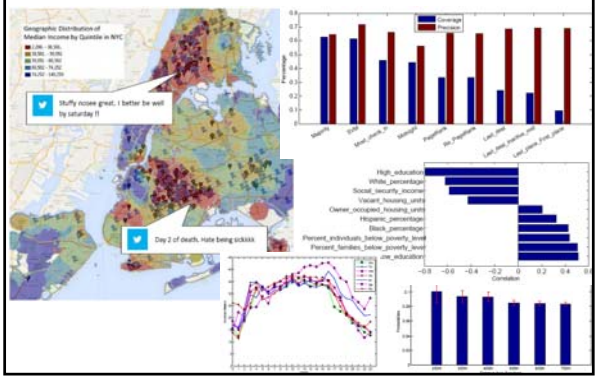
- New fraud schemes constantly "invented" by criminals
- Expensive to build (knowledge intensive)
- Difficult to maintain: frequently rules have to be updated



36

Home Location from Tweets & Urban Computing

Main Contributors: Tianran Hu, Henry Kautz, Jiebo Luo



37

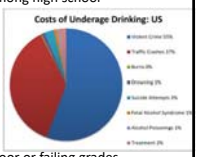
Using Social Multimedia to Solve Social Problems

Main Contributors: Ran Pang, Jiebo Luo, and Henry Kautz

Drinking Levels among Youth

The CDC 2011 Youth Risk Behavior Survey found that among high school students, during the past 30 days:

- 39% drank some amount of alcohol.
- 22% binge drank.
- 8% drove after drinking alcohol.
- 24% rode with a driver who had been drinking alcohol.



Consequences of Underage Drinking

- School problems, such as higher absence and poor or failing grades.
- Social problems, such as fighting, physical and sexual assault.
- Legal problems, such as arrest for driving or physically hurting someone while drunk.
- Physical problems, such as hangovers or illnesses.
- Unwanted, unplanned, and unprotected sexual activity.
- Higher risk for suicide and homicide.
- Alcohol-related car crashes and other unintentional injuries.
- Abuse of other drugs.

38

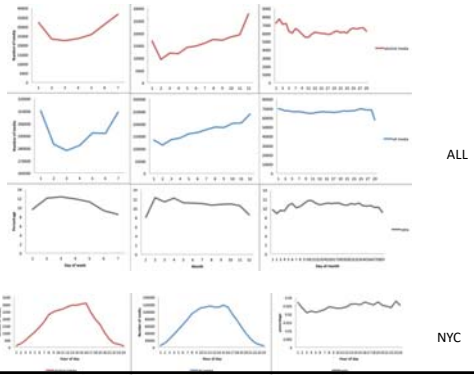
Using Social Multimedia to Solve Social Problems

Main Contributors: Ran Pang, Jiebo Luo, and Henry Kautz

39

Time Patten of Underage Alcohol Use

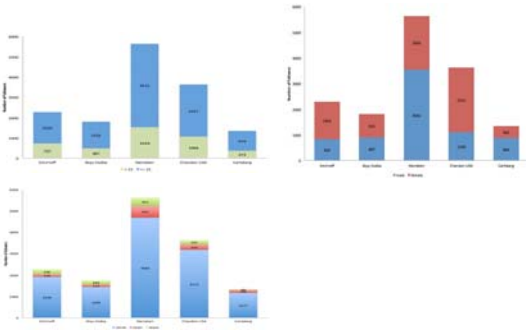
Main Contributors: Ran Pang, Jiebo Luo, and Henry Kautz



40

Brand Influence in Underage Alcohol Use

Main Contributors: Ran Pang, Jiebo Luo, and Henry Kautz



41

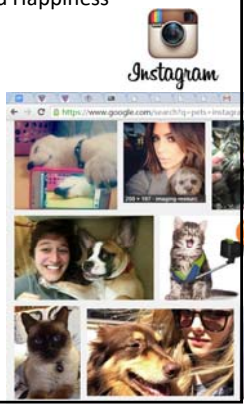
Understanding Pets and Happiness

Main Contributors: Yucheng Wu, Ran Pang, Jiebo Luo

Social support is critical for psychological and physical well-being, reflecting the centrality of belongingness in our lives. Human interactions often provide people with considerable social support, but can pets also fulfill one's social needs?

Studies found in a community sample that pet owners fared better on several well-being (e.g., greater self-esteem, more exercise) and individual-difference (e.g., greater conscientiousness, less fearful attachment) measures.

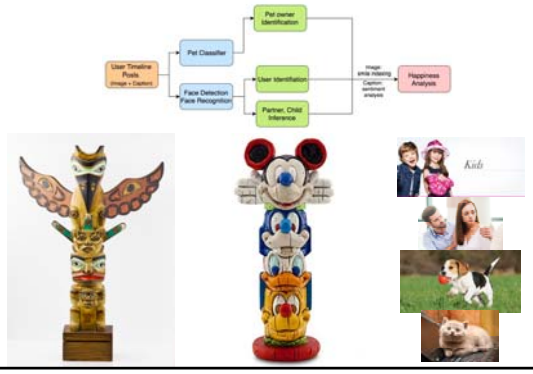
We intend to verify such findings at a larger scale and potentially at a fine granularity, through social multimedia. We will set up an *experimental group* and a *control group*.



42

The Totem Pole of Happiness

Main Contributors: Xuefeng Peng, Kevin Chi, Jiebo Luo



43

From Catwalk to Main Street

Main Contributors: Kezhen Chen, Kuan-Ting Chen*, Peizhong Cong, Winston Hsu, Jiebo Luo (MM '15 Grand Challenge)

Motivations

•In modern times, a growing number of people pay more attention to fashion and the mass has the penchant to emulate what large city residents and celebrities wear

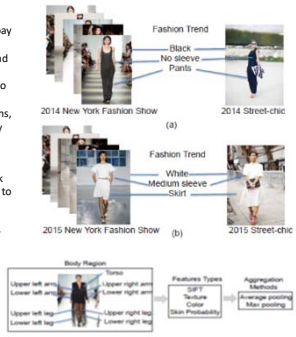
•Investigating fashion trends is of great interest to the industry and academia because of the potential for boosting many emerging applications, such as clothing recommendation, advertising by clothing brand association, etc.

Approach

1.Constructing a large dataset from the New York Fashion Shows and New York street chic in order to understand the likely clothing fashion trends in New York

2.Utilizing a learning-based approach to discover fashion attributes as the representative characteristics of fashion trends, and

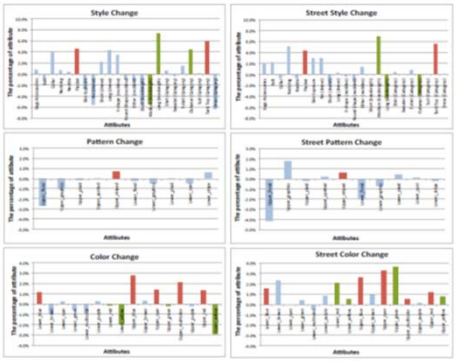
3.Comparing the analysis results from the New York Fashion Shows and street-chic images to verify whether the fashion shows have actual influence on the public in New York City.



44

From Catwalk to Main Street

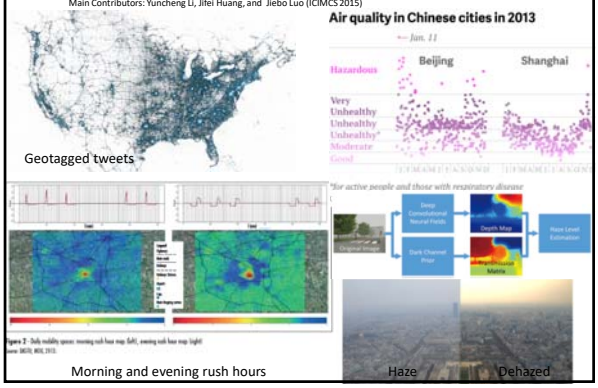
Main Contributors: Kezhen Chen, Kuan-Ting Chen*, Peizhong Cong, Jiebo Luo



45

Human Mobility and Human-Environment Interaction

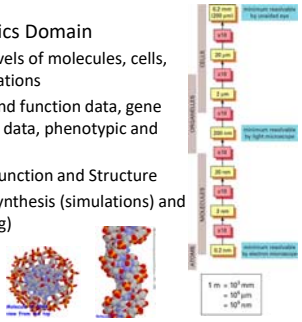
Main Contributors: Yuncheng Li, Jifei Huang, and Jiebo Luo (ICIMCS 2015)



46

Introduction: Microarray Problem in Bioinformatics

- Problems in Bioinformatics Domain
  - Data production at the levels of molecules, cells, organs, organisms, populations
  - Integration of structure and function data, gene expression data, pathway data, phenotypic and clinical data, …
  - Prediction of Molecular Function and Structure
  - Computational biology: synthesis (simulations) and analysis (machine learning)



47

47

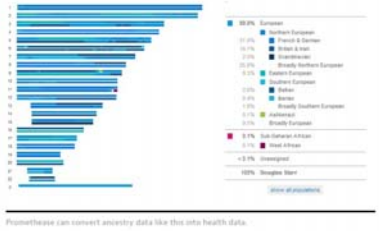
You Can Transform Your Genetic Ancestry Data Into Health Info, But Your Results May Vary

KQED Science NEWS PROGRAMS & BLOGS EDUCATION RESOURCES

By **Dr. Barry Starr**, Tech Museum & Stanford University

JUNE 30, 2014

PRINT SHARE



DNA ancestry tests like those offered by 23andMe or ancestry.com are a

48

Data Mining for Various Applications

- Mining data streams: Tilted time-window, micro-clustering, biased sampling & ensemble
- Mining spatial and multimedia data: Progressive deepening, co-location, invariant analysis
- Mining moving objects, trajectories, RFID/sensor networks
  - Sub-trajectory partitioning and regrouping
- Mining text and Web: PageRank to Web community discovery, opinion & usage mining, veracity analysis, …
- Mining software bugs, system performance data
- Mining biological data
- Privacy-preserving data mining
- Invisible data mining

49

Textbook

- Required: "Data Mining: Concepts and Techniques", 3/E, by Jiawei Han, Michelin Kamber and Jian Pei. All the homework assignments are from this book.
- Optional: "Mining of Massive Datasets", 2/E, Jure Leskovec, Anand Rajaraman, Jeffery David Ullman. More advanced.
- Course webpage: through my URCS page, or Blackboard (up to date)

50

Syllabus/Topics (2019)

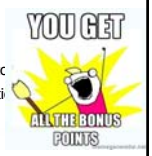
Overview and Introduction	(notes, Chap. 1)
Getting to Know Your Data	(Chap. 2)
Data Preprocessing	(Chap. 3)
Linear Algebra, Statistics Review	(notes)
Pattern Recognition Concepts	(notes, Duda & Hart)
Mining Frequent Patterns	(Chap. 6)
Association and Correlation	(Chap. 6)
Advanced Pattern Mining	(Chap. 7)
Classification	(Chap. 8/9*)
Cluster Analysis	(Chap. 10/11*)
Outlier detection	(Chap. 12)
Advanced Topics: Social Media Mining	(Special Lecture)
Advanced Topics: Bioinformatics	(Guest Lecture: Martin Zand/Tim Dye)
Advanced Topics: Network Mining*	(Guest Lecture*: Gourab Ghoshal)
Trends and Research Frontiers	(Chap. 13, notes)

- * time permitting

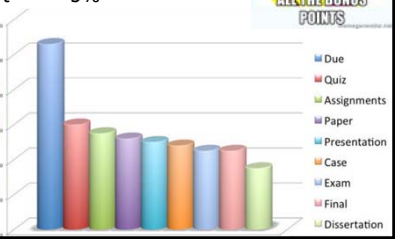
51

Grading

Homework	25% (5 x 5)
Mid-term	30% (close book)
Project I	10% (same assignment to
Project II	30% (including presentati
Attendance/effort	5%



COURSE POLICY:
- *Course Prerequisites:* M Python/Java programmin
- *Homeworks and Projec* **points** will be subtract submit in **Blackboard**
- *Exam policy*: No make- a valid reason, the projec



54

Course Project

- Something 'NEW' is expected
  - New problem, new method
  - Old problem, new method
  - Old method, new problem
  - New comparative study

Who needs papers? **You** do!

- Setting a high target: a conference paper
- 2-person teams (both must present!)
  - Project proposal
    - Problem statement & related work
    - Data source
    - Software tools and/or programming language(s)
    - Main references
  - Project presentation (1 in 2 selected per *interestingness*)

55

Team Project?

56

More on course projects

- Your own idea
- Suggested topics

57

What Can You Expect to Accomplish?

- General knowledge of the field
- Major concepts/approaches
- Exposure to extensive real-world problem-solving examples
- Hands-on experience with solving a real-world data mining problem
- Appreciation of data and data mining
- Awareness of the field: What's hot today? Where is DM going?
- Inspiration – Where can you use DM?

58

What You CANNOT Expect

- Data science is for *everyone*
- An easy course with *light* materials [to get B or better]
- ***Survival*** without *programming* skills (CSC 171/172)
- ***Success*** without general knowledge of AI (e.g. CSC 242), math (CSC 161/164) and probability (e.g. CSC 262)
- Doing well while skipping classes and spending little time on textbook
- Getting good grades on assignments by working only the night before the due date
- Giving practice questions for midterm exam (would your future boss do that? He does not even know!)
- A lot of hand-holding as in typical undergraduate courses (This is an advanced course; I will NOT spend one lecture going through one equation: I WILL help you though)
- Getting an "A" easily (>30% students DID get A/A-)
- Whining about any of the above in the course evaluation (No, undergraduates do NOT have any disadvantages)

59

Teaching Philosophy

- I will teach a 400-level course
- Not 200-level (grow up, or wait until grown up) 6
- For those taking 240: if you meet the same requirements, you get a bonus of 5%; if not quite, you get a prorated bonus.
- Certainly not 100-level
- Can anyone explain the differences?
- How do you know if you are ready for it?
  - Prerequisites: B or better
  - Freshman/sophomore (what's the hurry?), non-CS: placement test
  - "What if I'm not ready but required to take it now"
- Are you up to the challenge?

60

Testimonials

- "It gives students the freedom to study whatever sub-topics they are most interested in. I spent most of my time studying Bioinformatics, but other students spent time studying card games. There is a high level of freedom. "

62

Testimonials

Ulrik Soderstrom '16, '17 (MS) is one of the first Rochester students to graduate with a BA in data science.

Soderstrom is interested in the intersection of data science and environmental sustainability.

- **What's your favorite class that you've taken at the University of Rochester?**
- "Last semester, Data Mining with Jiebo Luo was an excellent course. I think he does a good job of giving students a range of skills that go just in-depth enough and then allowing them a creative space for an open-ended project at the end. I worked on a data mining project that determined the effects of industry-specific events—such as the impact of the presidential debates—on stock differences of companies like General Motors or Exxon. That's really rare in a lot of academics to get free rein like that and be responsible for your project."

63

A (sometimes) dreaded question

64

A sometimes (dreaded) question

~~C/C++~~

Java

python

65

Making This Course Work

- Working together
  - We have a very diverse, very large class
  - We will cover a lot of materials
  - Your cooperation: assignments on time, study on your own
  - You are here to learn, not just to get a grade
- Lectures: Powerpoint (occasionally board)
  - ***You*** should take notes
  - ***You*** should read the textbook (*before/after* the lecture)
- Course Schedule
  - Guest lectures
  - Instructor's office hours (T/R 3-4pm)
  - Up to 4 TAs: Tianlang Chen (M/W 3:30-4:30pm), Numair Sani (M/W 1-2pm), Yiming Pan (T/R 2-3pm) in WEGMANS 3504
- Communication
  - *Blackboard* (only using emails when necessary)

66

An Alternative

- Business Analytics (offered by Simon School)
  - Lighter treatment
  - Spreadsheet/R
- Do what's right for YOU, ignore the HYPE

67

Academic Honesty

- You are responsible for knowing the College of Arts, Sciences and Engineering and course policies on academic honesty. Read your syllabus for the academic honesty policy for this course. Talk to me if you have any questions.
- I take violations of academic honesty seriously. Suspected violations will be pursued vigorously following the College's procedures for suspected cases of academic dishonesty (see URL link below)

Some Categories of Academic Dishonesty

- **Plagiarism:** Representing someone else's work as your own
- **Cheating**: Using unauthorized information or sources for an assignment or exam.
- **Assisting others in academic dishonesty**
- **Falsifying Information**
- **Interfering with others' access to legitimate course materials**

This is not an all-inclusive list. For more information, consult with your instructor, the student handbook and visit the URL below.

http://www.rochester.edu/College/Honesty

68

Academic Honesty

Summarized by George Ferguson

- Responsibilities of instructors: http://www.rochester.edu/college/honesty/instructors.html
  Include an academic honesty statement on each course syllabus or the course Blackboard page or website about how academic honesty applies to the course, and call attention to the information during at least one class session during the first two weeks of class." [continues...]
- On the question of graduate students being covered by the Academic Honesty policy: http://www.rochester.edu/college/honesty/returninggraduates.html
  "Graduate students in Arts, Sciences, and Engineering are now included under the College Academic Honesty Policy. Before, if a graduate student were accused of academic dishonesty, the case would be resolved within their department before going to the Graduate Dean of AS&E. To improve consistency and increase fairness, if a graduate student is suspected of academic dishonesty in their coursework,it will now be resolved through Board on Academic Honesty processes."
- Note also that the excuse of "inexperience" is not permitted for graduate students: http://www.rochester.edu/college/honesty/instructors.html#suspect
- On the question of the honor pledge in quizzes, the wording says "all examinations": http://www.rochester.edu/college/honesty/policy.html#pledge
  "The following Honor Pledge will be copied and signed by all students on all examinations: "I affirm that I will not give or receive any unauthorized help on this exam, and that all work will be my own."

69

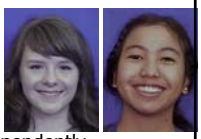
From a Computer Science course

- **2) A model for courses that involve collaborative work in laboratory or problem sets, :**
- Academic honesty: www.rochester.edu/college/honesty/ www.rochester.edu/college/CCAS/AdviserHandbook/AcadHonesty.html *Homework/Project collaboration:* **You may discuss homework problems with others**, but you must **not retain** written notes from your conversations with other students, or share data via computer files to be used in completing your homework. Your written work must be completed without reference to such notes, with the exception of class and recitation notes, which may be retained in written form. [NOTE: some instructors require students to report the names of those with whom they discussed an assignment.]
- *General rule:* When in doubt, cite (algorithm, code, data from online)

Aptitude/Placement Test

- Absolute basics
  - Freshman/Sophomore, non-CS
  - Quantitative
  - Programming
  - Honesty
  - Meeting deadlines

Midterm: October 26 or 31
Project Presentation:
11/30, 12/5, 12/7, 12/12

Want to do well in 240/440?

- Absolutely no plagiarism
- Follow my instructions (religiously)
  - Meet the prerequisites!
  - Take notes
  - Read the textbook
  - Work on assignments early and independently
  - Make use of office hours
  - Prepare for midterm
  - Put effort in the course project
  - Forget about whining!
    - If you claim you meet the prerequisites
    - If you don't turn in assignments
    - If you skip classes w/o a good reason
    - …